On high-order multilevel optimization strategies*

Henri Calandra†, Serge Gratton‡, Elisa Riccietti‡, and Xavier Vasseur§

Abstract. We propose a new family of multilevel methods for unconstrained minimization. The resulting strategies are multilevel extensions of high-order optimization methods based on q-th order Taylor models (with $q \geq 1$) that have been recently proposed in the literature. The use of high-order models, while decreasing the worst-case complexity bound, makes these methods computationally more expensive. Hence, to counteract this effect, we propose a multilevel strategy that exploits a hierarchy of problems of decreasing dimension, still approximating the original one, to reduce the global cost of the step computation. A theoretical analysis of the family of methods is proposed. Specifically, local and global convergence results are proved and a worst-case complexity bound to reach first-order stationary points is also derived. A multilevel version of the well-known adaptive regularization by cubics (corresponding to $q = 2$ in our setting) has been implemented, as well as a multilevel third-order method ($q = 3$). Numerical experiments clearly highlight the relevance of the new multilevel approaches leading to considerable computational savings compared to their one-level counterparts.

Key words. nonlinear optimization, multilevel methods, high-order optimization model, global convergence, complexity estimation, error bound, tensor methods.

AMS subject classifications. 90C30, 65K05, 90C26, 90C06

1. Introduction.

We propose a new family of high-order multilevel optimization methods for unconstrained minimization. Exploiting ideas stemming from multilevel methods allows us to reduce the cost of the step computation, which represents the major cost per iteration of the standard single level procedures. We have been mainly inspired by two driving ideas: the use of high-order models in optimization as introduced in [6], and the multilevel recursive strategy proposed in [21].

When solving unconstrained minimization problems, quadratic models are widely used. These are usually regularized by a quadratic term. For example, trust-region methods have been widely studied and used to globalize Newton-like iterations [18, 38]. Lately in the literature, a different option has received a growing attention: the use of a cubic overestimator of the objective function as a regularization technique for the computation of the step from one iterate to the next, giving rise to quadratic models with cubic regularization. This idea first appeared in [22] and then was reconsidered in [37], where the authors proved that the method has a better worst-case complexity bound compared to standard trust-region methods. Later, in [15, 16], an adaptive variant of the method has been proposed, based on a dynamic choice

*Submitted to the editors March 2, 2020.

Funding: This work was funded by TOTAL.

†TOTAL, Centre Scientifique et Technique Jean Féger, avenue de Larribau F-64000 Pau, France (henri.calandra@total.com).

‡INPT-IRIT, University of Toulouse and ENSEEIHT, 2 Rue Camichel, BP 7122, F-31071 Toulouse Cedex 7, France (serge.gratton@enseeiht.fr, elisa.riccietti@enseeiht.fr).

§ISAE-SUPAERO, University of Toulouse, 10, avenue Edouard Belin, BP 54032, F-31055 Toulouse Cedex 4, France (xavier.vasseur@isae-supaero.fr).

of the regularization parameters and on an approximate solution of the subproblems. The resulting method is known as adaptive method based on cubic regularization (ARC) and is shown to preserve the attractive global complexity bound established in [37]. In recent years the method has attracted further interest, see for example [14, 20, 40, 43].

In recent publications also methods of higher order start to gain interest, see for example [6, 41]. In [6] in particular, it has been observed that the good complexity bound of ARC can be made even lower, if one is willing to use higher-order derivatives. In specific applications this computation is indeed feasible, for example when considering partially separable functions [17]. The authors in [6] present a family of methods based on high-order regularized models, which are a generalization of the second-order model employed in ARC. Specifically, they are based on models of order $q \geq 1$, regularized by a term of order $q+1$. The method based on cubic regularization belongs to this family and corresponds to the choice $q = 2$. The authors in [6] propose a unifying framework to describe the theoretical properties of the approaches in this class. It is proved that the method based on the q-th order model requires at most $O\left(\epsilon^{-\frac{q+1}{q}}\right)$ function evaluations to find a first-order critical point, where ϵ denotes the absolute accuracy level.

However, the use of higher-order models comes along with higher computational costs. The main cost per iteration of the methods described in [6] is represented by the step computation through the model minimization. This cost is proportional to the dimension of the problem, it can therefore be significant for large-scale problems. This issue has been faced in the literature by extending to nonlinear optimization ideas coming from multigrid [25], giving rise to a wide variety of methods to solve both convex [3, 4, 19, 24, 28, 29, 30, 33, 39] and nonconvex problems [21, 26, 31, 32, 35, 36, 42].

These methods share with classical multigrid methods the idea of exploiting a hierarchy of problems (in this case a sequence of nonlinear functions) defined on lower dimensional spaces, approximating the original objective function f. The simplified expressions of the objective function are used to build models that are cheaper to minimize, and are used to define the step. Specifically in [21] the authors present an extension of classical multigrid methods for nonlinear optimization problems (see [8, 9] or [11, Ch. 3]) to a class of multilevel trust-region based optimization algorithms.

Multilevel approaches can be used in every situation in which a hierarchy of functions defined on increasingly lower dimensional spaces is available. This is the typical scenario when the problem arises from the discretization of an infinite dimensional problem and increasingly coarser discretizations can be used to define the hierarchy. There are however also examples of discrete problems that can be provided with a hierarchy built exclusively by algebraic procedures, see for example [12].

Our contributions. Inspired by the ideas presented in [6, 21] we propose a family of multilevel optimization methods using high-order regularized models that generalizes the methods proposed in both papers. The aim is to decrease the computational cost of the methods in [6] extending the ideas in [21] to higher-order models. We also develop a theoretical analysis for the resulting family of methods. The main theoretical results are provided in Theorems 4.2, 4.3 and 4.7, respectively. In these theorems we successively prove the global convergence

property of the methods, evaluate a worst-case complexity bound to reach a first-order critical point and provide local convergence rates. The global convergence analysis generalizes the results in [6] and appears as much simpler than that in [21]. Moreover we establish local convergence results towards second-order stationary points, that are not present neither in [6] nor in [21]. These results not only generalize those in [44], that are valid only for $q = 2$, but also apply to the one level methods in [6]. From a practical point of view, we implemented the methods of the family corresponding to $q = 2$ (which represents a multilevel version of the well-known adaptive regularization by cubics) and to $q = 3$.

To the best of our knowledge, this is the first time that multilevel optimization strategies, based on models of generic order $q \geq 1$, are proposed, and that a unifying framework is introduced to study their convergence. In particular, multilevel versions of the adaptive regularization by cubics have never been analysed nor tested numerically before. Moreover, this is the first time that local convergence results are proposed for q-th order methods.

Structure. The manuscript is organized as follows. In Section 2, we briefly introduce the family of optimization methods using high-order regularized models considered in [6]. Section 3 and Section 4 represent our main contribution. We introduce in Section 3 the multilevel extensions of the methods presented in Section 2, and we provide a theoretical analysis in Section 4. Specifically, we focus on global convergence in Section 4.1, worst-case complexity in Section 4.2 and local convergence in Section 4.3. In Section 5 we then present results related to numerical experiments performed with the multilevel methods corresponding to $q = 2, 3$. Finally, conclusions are drawn in Section 6.

Tensor notations. To deal with high-order derivatives we will need to use a tensor notation, that we introduce here for convenience of the reader, see [6, 7]. We first consider a tensor of order three, and then extend the definition to a tensor of order $p \in \mathbb{N}$.

Definition 1.1. *Let $T \in \mathbb{R}^{n\times n\times n}$, and $u, v, w \in \mathbb{R}^n$. Then $T[u, v, w] \in \mathbb{R}$, $T[u, v] \in \mathbb{R}^n$ and*

$$T[u,v,w] = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} T(i,j,k)u(i)v(j)w(k),$$
$$T[v,w](i) = \sum_{j=1}^{n}\sum_{k=1}^{n} T(i,j,k)v(j)w(k), \quad i = 1,\dots,n.$$

Definition 1.2. *Let $p \in \mathbb{N}$ and $T \in \mathbb{R}^{n^p}$, and $u_1, \dots, u_p \in \mathbb{R}^n$. Then $T[u_1, \dots, u_p] \in \mathbb{R}$, $T[u_1, \dots, u_{p-1}] \in \mathbb{R}^n$ and*

$$T[u_1,\dots,u_p] = \sum_{j_1=1}^{n}\cdots\sum_{j_p=1}^{n} T(j_1,\dots,j_p)u_1(j_1)\dots u_i(j_p),$$
$$T[u_1,\dots,u_{p-1}](j_1) = \sum_{j_2=1}^{n}\cdots\sum_{j_p=1}^{n} T(j_1,\dots,j_p)u_1(j_2),\dots u_{p-1}(j_p), \quad j_1 = 1,\dots,n.$$

More generally, for a tensor T of order p, $T[u_1, \dots, u_j]$ for $j \leq p$ is a tensor of order $p - j$ resulting from the application of T to the vectors $u_1, \dots, u_j$.

In the following we will use the notation $T\,[s]^i$, which stands for the tensor T applied i times to the vector $s \in \mathbb{R}^n$.

Given a function $f : \mathbb{R}^n \to \mathbb{R}$ we denote by $\nabla_x^p f(x)$ its p-th order derivative tensor

$$\nabla_x^p f(x) = \left[\frac{\partial^p f}{\partial x_{i_1} \dots \partial x_{i_p}} \right]_{i_j \in \{1,\dots,n\},\, j=1,\dots,p}.$$

For convenience, we will omit the superscript when $p = 1$.

We will denote by $\|\cdot\|$ the Euclidean norm and by $\|\cdot\|_{[p]}$ the tensor norm recursively induced by the Euclidean norm on the space of p-th order tensors, which for a tensor T of order p is given by

$$\|T\|_{[p]} := \max_{\|u_1\|=\dots=\|u_p\|=1} |T[u_1, \dots, u_p]|.$$

We will use the same notation for all the spaces we will consider, the space on which the norm is defined will be clear by the context.

2. High-order iterative optimization methods. Let $q \geq 1$ be an integer. Let us consider a minimization problem of the form:

$$\min_{x \in \mathbb{R}^n} f(x) \tag{2.1}$$

with $f : \mathbb{R}^n \to \mathbb{R}$ a bounded below and q-times continuously differentiable function, called the objective function.

Classical iterative optimization methods for unconstrained minimization are based on the use of a model to approximate the objective function at each iteration. In this section, we describe the iterative optimization methods using high-order models presented in [6].

2.1. Model definition and step acceptance. At each iteration k, given the current iterate x_k, the objective function is approximated by the q-th order Taylor series of f

$$T_{q,k}(x_k, s) = f(x) + \sum_{i=1}^{q} \frac{1}{i!} \nabla_x^i f(x)\,[s]^i. \tag{2.2}$$

A step s_k is then found minimizing (possibly approximately) the regularized Taylor model of order q:

$$m_{q,k}(x_k, s; \lambda_k) = T_{q,k}(x_k, s) + \frac{\lambda_k}{q+1} \|s\|^{q+1} \tag{2.3}$$

where λ_k is a positive value called regularization parameter. The step s_k is used to define a trial point i.e. $x_{k+1} = x_k + s_k$. At each iteration, it has to be decided whether to accept the step or not. This decision is based on the accordance between the decrease in the function and in the model. More precisely, at each iteration both the decrease achieved in the model, that we call *predicted reduction*, $pred = T_{q,k}(x_k, 0) - T_{q,k}(x_k, s_k)$, and that achieved in the

objective function, that we call *actual reduction*, $ared = f(x_k) - f(x_k + s_k)$, are computed. The step acceptance is then based on the ratio:

$$\rho_k = \frac{ared}{pred} = \frac{f(x_k) - f(x_k + s_k)}{T_{q,k}(x_k, 0) - T_{q,k}(x_k, s_k)}. \tag{2.4}$$

If the model is accurate, ρ_k will be close to one. Then, the step s_k is accepted if ρ_k is larger than or equal to a chosen threshold $\eta_1 \in (0, 1)$ and is rejected otherwise. In the first case, the step is said to be *successful*, and otherwise the step is *unsuccessful*.

After the step acceptance, the regularization parameter is updated for the next iteration. The update is still based on the ratio (2.4). If the step is successful, the regularization parameter is decreased, otherwise it is increased. The whole procedure is stopped when a minimizer of f is reached. Usually, the stopping criterion is based on the norm of the gradient, i.e. given an absolute accuracy level $\epsilon > 0$ the iterations are stopped as soon as $\|\nabla_x f(x_k)\| < \epsilon$.

2.2. Minimization of the model. The main computational work per iteration in this kind of methods is represented by the minimization of the regularized model (2.3), whose cost naturally depends on the dimension of the problem. However, from the convergence theory of such methods, it is well known that it is not necessary to minimize the model exactly to get a globally convergent method.

A well-known possibility is to minimize the model until the Cauchy decrease is achieved, i.e. until a fraction of the decrease provided by the Cauchy step (the step that minimizes the model in the direction of the negative gradient) is obtained. In [6] the authors consider a different stopping criterion for the inner iterations:

$$m_{q,k}(x_k, s_k; \lambda_k) < m_{q,k}(x_k, 0; \lambda_k), \quad \|\nabla_s m_{q,k}(x_k, s_k; \lambda_k)\| \leq \theta \|s_k\|^q, \tag{2.5}$$

for $\theta > 0$, which has the advantage of allowing for simpler convergence proofs. The whole procedure is sketched in Algorithm 2.1.

For very large-scale problems however, even an approximate minimization of (2.3) may be really costly. Then, in the next section we propose multilevel variants of the procedures, that rely on simplified models of the objective function, cheaper to optimize, allowing to reduce the global cost of the optimization procedure.

3. Multilevel optimization methods. We describe the multilevel extension of the family of methods presented in Section 2. The procedures are inspired by the multilevel trust-region approach presented in [21], where only second-order models with quadratic regularization have been considered. Here, we generalize this approach by allowing also higher-order models, i.e. $q > 2$.

3.1. Preliminaries and notations. In standard optimization methods the minimization of (2.3) represents the major cost per iteration, which crucially depends on the dimension n of the problem. When n is large, the solution cost is therefore often significant. We want to reduce this cost by exploiting the knowledge of alternative simplified expressions of the objective function. More specifically, we assume that we know a collection of functions $\{f_l\}_{l=1}^{l_{\max}}$ such that each f_l is a q-times continuously differentiable function from $\mathbb{R}^{n_l} \to \mathbb{R}$

Algorithm 2.1 AR**q**$(x_0, \lambda_0, \epsilon)$ (Adaptive Regularization method of order q)

Given $0 < \eta_1 \leq \eta_2 < 1$, $0 < \gamma_2 \leq \gamma_1 < 1 < \gamma_3$, $\lambda_{\min} > 0$, $\theta > 0$.
Input: $x_0 \in \mathbb{R}^n$, $\lambda_0 > \lambda_{\min}$, $\epsilon > 0$.
$k = 0$
while $\|\nabla_x f(x_k)\| > \epsilon$ **do**
  • **Initialization:** Define the model $m_{q,k}$ as in (2.3).
  • **Model minimization:** Find a step s_k that sufficiently reduces the model $m_{q,k}$, i.e. that satisfies (2.5) .
  • **Acceptance of the trial point:** Compute $\rho_k = \dfrac{f(x_k) - f(x_k + s_k)}{T_{q,k}(x_k, 0) - T_{q,k}(x_k, s_k)}$.
  if $\rho_k \geq \eta_1$ **then**
    $x_{k+1} = x_k + s_k$
  else
    $x_{k+1} = x_k$.
  end if
  • **Regularization parameter update:**
  if $\rho_k \geq \eta_1$ **then**

$$\lambda_{k+1} = \begin{cases} \max\{\lambda_{\min}, \gamma_2 \lambda_k\}, & \text{if } \rho_k \geq \eta_2, \\ \max\{\lambda_{\min}, \gamma_1 \lambda_k\}, & \text{if } \rho_k < \eta_2, \end{cases}$$

  else
    $\lambda_{k+1} = \gamma_3 \lambda_k$.
  end if
  $k = k + 1$
end while

and $f^{l_{\max}}(x) = f(x)$ for all $x \in \mathbb{R}^n$. We will also assume that, for each $l = 2, \ldots, l_{\max}$, f_l is more costly to minimize than f_{l-1}. This is the typical scenario when the problem arises from the discretization of an infinite dimensional problem, the f_l's represent increasingly finer discretizations and $n_l \geq n_{l-1}$ for all l. This is of course not the only possible application. For example, an interesting scenario in which no discretization is involved, arises in the training of artificial neural networks. In this case a multilevel algorithm can be used for the training and a hierarchy of networks with less and less neurons can be built (even if there is no underlying geometrical structure) for example by means of algebraic multigrid strategies [12]. As we do not assume the hierarchy to come from a discretization process, we do not use the terminology typically used in the field of multigrid methods, and we use 'levels' rather than 'grids'.

The methods we propose are recursive procedures, so it suffices to describe the two-level case. For sake of simplicity from now on we will assume that we have just two approximations to our objective f at disposal. This amounts to consider $l_{\max} = 2$. For ease of notation, we will denote by $f^h : \mathbb{R}^{n_h} \to \mathbb{R}$ the approximation at the highest level ($f^h(x) = f^{l_{\max}}(x)$ in the notation previously used) and by $f^H : \mathbb{R}^{n_H} \to \mathbb{R}$ the other approximation available, that is cheaper to optimize. The quantities on the highest level will be denoted by a superscript h, whereas the quantities on the lower level will be denoted by a superscript H. Let x_k^h denote

the k-th iteration at the highest level.

3.2. Construction of the lower level model. The main idea is to use f^H to construct, in the neighborhood of the current iterate, an alternative model $t^H_{q,k}$ to the Taylor model $T^h_{q,k}$ in (2.2) for $f^h = f$ [21]. The alternative model $t^H_{q,k}$ should be cheaper to optimize than $T^h_{q,k}$, and will be used, whenever suitable, to define the step. Of course, for f^H to be useful at all in minimizing f^h, there should be some relation between the variables of these two functions. We henceforth assume the following.

Assumption 1. *Let us assume that there exist two full-rank linear operators $R : \mathbb{R}^{n_h} \to \mathbb{R}^{n_H}$ and $P : \mathbb{R}^{n_H} \to \mathbb{R}^{n_h}$ such that $P = \alpha R^T$, for a fixed scalar $\alpha > 0$. Let us assume also that it exists $\kappa_R > 0$ such that $\max\{\|R\|, \|P\|\} \leq \kappa_R$, where $\|\cdot\|$ denotes the matrix norm induced by the Euclidean norm at the fine level.*

In the following, we assume $\alpha = 1$ without loss of generality, as the problem can be easily scaled to handle the case $\alpha \neq 1$.

At each iteration k at highest level we set $x^H_{0,k} = R\, x^h_k$, i.e. the initial iterate at the lower level is set as the projection of the current iterate, and we define the lower level model $t^H_{q,k}$ as a modification of the coarse function f^H. Given q, f^H is modified adding q correction terms, to enforce the following relation:

$$\nabla^i_s t^H_{q,k}(x^H_{0,k}) \left[s^H\right]^i = \mathcal{R}(\nabla^i_x f^h(x^h_k)) \left[s^H\right]^i, \quad i = 1, \ldots, q, \tag{3.1}$$

where $\mathcal{R}(\nabla^i_x f^h(x^h_k))$ is such that for all $i = 1, \ldots, q$ and $s^H_1, \ldots, s^H_i \in \mathbb{R}^{n_H}$

$$\begin{aligned} \mathcal{R}(\nabla^i_x f^h(x^h_k))[s^H_1, \ldots, s^H_i] &:= \nabla^i_x f^h(x^h_k)[Ps^H_1, \ldots, Ps^H_i], \\ \langle \mathcal{R}(\nabla^i_x f^h(x^h_k))[s^H_1, \ldots, s^H_{i-1}], s^H_i \rangle &:= \langle \nabla^i_x f^h(x^h_k)[Ps^H_1, \ldots, Ps^H_{i-1}], Ps^H_i \rangle, \end{aligned} \tag{3.2}$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product.

For instance, if $q = 2$, relation (3.1) simply becomes:

$$\nabla_s t^H_{q,k}(x^H_{0,k})^T s^H = (R\ \nabla_x f^h(x^h_k))^T s^H, \quad (s^H)^T \nabla^2_x t^H_{q,k}(x^H_{0,k}) s^H = (s^H)^T R\ \nabla^2_x f^h(x^h_k)\ Ps^H.$$

Relation (3.1) crucially ensures that the behaviours of f^h and $t^H_{q,k}$ are coherent up to order q in a neighbourhood of x^h_k and $x^H_{0,k}$. To achieve (3.1), we define the lower level model $t^H_{q,k}$ as

$$t^H_{q,k}(x^H_{0,k}, s^H) = f^H(x^H_{0,k} + s^H) + \sum_{i=1}^{q} \frac{1}{i!} \left(\mathcal{R}(\nabla^i_x f^h(x^h_k)) - \nabla^i_x f^H(x^H_{0,k}) \right) \left[s^H\right]^i, \tag{3.3}$$

with $\mathcal{R}(\nabla^i_x f^h(x^h_k))$ defined in (3.2). When $q = 2$ this is simply:

$$\begin{aligned} t^H_{2,k}(x^H_{0,k}, s^H) =& f^H(x^H_{0,k} + s^H) + (R\nabla_x f^h(x^h_k) - \nabla_x f^H(x^H_{0,k}))^T s^H \\ &+ \frac{1}{2}(s^H)^T (R\nabla^2_x f^h(x^h_k)P - \nabla^2_x f^H(x^H_{0,k})) s^H. \end{aligned}$$

3.3. Step computation and step acceptance. At each generic iteration k of our method, a step s_k^h has to be computed to define the new iterate. Then, one has the choice between the Taylor model (2.2) and a lower level model (3.3).

Obviously, it is not always possible to use the lower level model. For example, it may happen that $\nabla_x f^h(x_k^h)$ lies in the nullspace of R and thus that $R\nabla_x f^h(x_k^h)$ is zero while $\nabla_x f^h(x_k^h)$ is not. In this case, the current iterate appears to be first-order critical for $t_{q,k}^H$ while it is not for f^h. Using the model $t_{q,k}^H$ is hence potentially useful only if $\|\nabla_s t_{q,k}^H(x_{0,k}^H)\| = \|R\nabla_x f^h(x_k^h)\|$ is large enough compared to $\|\nabla_x f^h(x_k^h)\|$ [21]. We therefore restrict the use of the model $t_{q,k}^H$ to iterations where

$$\|R\nabla_x f^h(x_k^h)\| \geq \kappa_H \|\nabla_x f^h(x_k^h)\| \quad \text{and} \quad \|R\nabla_x f^h(x_k^h)\| > \epsilon_H, \tag{3.4}$$

for some constant $\kappa_H \in (0, \min\{1, \|R\|\})$ and where $\epsilon_H \in (0,1)$ is a measure of the first-order criticality for $t_{q,k}^H$ that is judged sufficient at level H [21]. Note that, given $\nabla_x f^h(x_k^h)$ and R, this condition is easy to check before even attempting to compute a step at a lower level.

If the Taylor model is chosen, then we just compute a step as in standard methods, minimizing (possibly approximately) the corresponding regularized model

$$m_{q,k}^h(x_k^h, s^h; \lambda_k) = T_{q,k}^h(x_k^h, s^h) + \frac{\lambda_k}{q+1}\|s^h\|^{q+1}, \tag{3.5}$$

with $T_{q,k}^h$ the Taylor series of f^h as defined in (2.2). If the lower level model is chosen, we then minimize (possibly approximately) the following regularized model:

$$m_{q,k}^H(x_{0,k}^H, s^H; \lambda_k) = t_{q,k}^H(x_{0,k}^H, s^H) + \frac{\lambda_k}{q+1}\|s^H\|^{q+1} \tag{3.6}$$

and obtain a point $x_{*,k}^H$ such that (if the minimization is successful) the value of the regularized model has been reduced, and a step $s_k^H = x_{*,k}^H - x_{0,k}^H$ (note that the iteration indices always refer to the highest level, we are not indexing the iterations on the lower level for the minimization of the lower level model). This step has to be prolongated back on the fine level, i.e. we define $s_k^h = Ps_k^H$. The minimization process is stopped as soon as the stopping condition

$$m_{q,k}^l(x^l, s^l; \lambda_k) < m_{q,k}^l(x^l, 0; \lambda_k) \text{ and } \|\nabla_s m_{q,k}^l(x^l, s^l; \lambda_k)\| \leq \theta_k^l \|s^l\|^q, \tag{3.7}$$
$$(x^l, s^l) = \begin{cases} (x_k^h, s_k^h) & \text{if} \quad l = h, \\ (x_{0,k}^H, s_k^H) & \text{if} \quad l = H, \end{cases}$$

is satisfied, for $l = h$ and $l = H$, respectively, where $\{\theta_k^h, \theta_k^H\}$ are bounded from above sequences such that $\theta_k^h, \theta_k^H \leq \theta_{\max}$ for all k. Note that in [6] a constant θ is considered. In both cases we are sure that it will exist a point that satisfies (3.7), as when the level is selected, a standard one-level optimization method is used, and the analysis in [6] applies.

In both cases, after the step is found, we have to decide whether to accept it or not. The step acceptance is based on the ratio:

$$\rho_k = \frac{f^h(x_k^h) - f^h(x_k^h + s_k^h)}{t_{q,k}^h(x_k^h, 0) - t_{q,k}^h(x_k^h, s_k^h)},$$

where

$$t_{q,k}^h(x_k^h, s^h) = \begin{cases} T_{q,k}^h(x_k^h, s^h) & \text{(Taylor model)}, \\ t_{q,k}^H(Rx_k^h, s^H),\ s^h = Ps^H & \text{(lower level model)}. \end{cases} \tag{3.8}$$

As in the standard form of the methods, the step is accepted if it provides a sufficient decrease in the function, i.e. if given $\eta_1 > 0$, $\rho_k \geq \eta_1$. The regularization parameter is also updated as in Algorithm 2.1. We sketch the whole procedure in Algorithm 3.1.

Some comments are necessary to explain Step 6 in Algorithm 3.1. The generic framework sketched in Algorithm 3.1 comprises different possible methods. Specifically, one of the flexible features (inherited by the method in [21]) is that, to ensure convergence, the minimization at lower levels can be stopped after the first successful iteration, as we will see in the next section. This therefore opens the possibility to consider both fixed form recursion patterns and free form ones. A free form pattern is obtained when Algorithm 3.1 is run carrying the minimization at each level out, until the norm of the gradient becomes small enough. The actual recursion pattern is then uniquely determined by the progress of minimization at each level and may be difficult to forecast. By contrast, the fixed form recursion patterns are obtained by specifying a maximum number of successful iterations at each level, a technique directly inspired from the definitions of V- and W-cycles in multigrid algorithms [25].

4. Convergence theory. In this section, we provide a theoretical analysis of the proposed family of multilevel methods. Inspired by the convergence theory reported in [6], we prove global convergence of the proposed methods to first-order critical points and we provide a worst-case complexity bound to reach such a point, generalizing the theory proposed in [6, 21]. At the same time the proposed analysis also appears as simpler than that in [21], since the regularization parameter λ_k is directly updated, rather than the trust-region radius, and since we use the stopping criterion (3.7). Moreover, we also propose local convergence results, which also apply to the methods in [6], and that extend those in [44] to higher-order models.

Note that, as the methods are recursive, we can restrict the analysis to the two-level case. For the analysis we need the following regularity assumptions as in [6].

Assumption 2. *Let f^h and f^H be q-times continuously differentiable and bounded below functions. Let us assume that the q-th derivative tensors of f^h and f^H are Lipschitz continuous, i.e. that there exist constants L_h, L_H such that*

$$\|\nabla_x^q f^l(x) - \nabla_x^q f^l(y)\|_{[q]} \leq (q-1)!\, L_l \|x - y\| \quad \text{for all} \quad x, y \in \mathbb{R}^{n_l}, \quad l = h, H.$$

We remind three useful relations, following from Taylor's theorem, see for example relations (2.3) and (2.4) in [6].

Lemma 4.1. *Let $g : \mathbb{R}^n \to \mathbb{R}$ be a q-times continuously differentiable function with Lipschitz continuous q-th order tensor, with L the corresponding Lipschitz constant. Given its q-th order*

Algorithm 3.1 MAR**q**$(l, f^l, x_0^l, \lambda_0^l, \epsilon^l)$ (Multilevel Adaptive Regularization method of order q)

Input: $l \in \mathbb{N}$ (index of the current level, $1 \leq l \leq l_{\max}$, $l_{\max}$ being the highest level), $f^l : \mathbb{R}^{n_l} \to \mathbb{R}$ function to be optimized ($f^{l_{\max}} = f$), $x_0^l \in \mathbb{R}^{n_l}$, $\lambda_0^l > \lambda_{\min}$, $\epsilon^l > 0$.
Given $0 < \eta_1 \leq \eta_2 < 1$, $0 < \gamma_2 \leq \gamma_1 < 1 < \gamma_3$, $\lambda_{\min} > 0$.
R_l denotes the restriction operator from level l to $l-1$, P_l the prolongation operator from level $l-1$ to l.
$k = 0$
while $\|\nabla_x f^l(x_k^l)\| > \epsilon^l$ **do**
• **Model choice:** If $l > 1$ compute $R_l \nabla_x f^l(x_k^l)$ and check (3.4). If $l = 1$ or (3.4) fails, go to Step 7. Otherwise, choose to go to Step 7 or to Step 8.
• **Taylor step computation:** Define $t_{q,k}^l(x_k^l, s^l) = T_{q,k}^l(x_k^l, s^l)$, the q-th order Taylor series of f^l. Find a step s_k^l such that (3.7) holds for $m_{q,k}^l(x_k^l, s^l; \lambda_k) = t_{q,k}^l(x_k^l, s^l) + \frac{\lambda_k}{q+1}\|s^l\|^{q+1}$. Go to Step 9.
• **Recursive step computation:** Define $x_{0,k}^{l-1} = R_l\, x_k^l$ and

$$t_{q,k}^{l-1}(x_{0,k}^{l-1}, s^{l-1}) = f^{l-1}(x_{0,k}^{l-1} + s^{l-1}) + \sum_{i=1}^{q} \frac{1}{i!}\left(\mathcal{R}(\nabla_x^i f^l(x_k^l)) - \nabla_x^i f^{l-1}(x_{0,k}^{l-1})\right)[s^{l-1}]^i,$$

$$m_{q,k}^{l-1}(x_{0,k}^{l-1}, s^{l-1}) = t_{q,k}^{l-1}(x_{0,k}^{l-1}, s^{l-1}) + \frac{\lambda_k}{q+1}\|s^{l-1}\|^{q+1}.$$

Choose ϵ^{l-1} and call MAR**q**$(l-1, m_{q,k}^{l-1}, x_{0,k}^{l-1}, \lambda_k^l, \epsilon^{l-1})$ yielding an approximate solution $x_{*,k}^{l-1}$ of the minimization of $m_{q,k}^{l-1}$ according to (3.7). Define $s_k^l = P_l\,(x_{*,k}^{l-1} - x_{0,k}^{l-1})$ and $t_{q,k}^l(x_k^l, s^l) = t_{q,k}^{l-1}(x_{0,k}^{l-1}, s^{l-1})$ for all $s^l = P s^{l-1}$.
• **Acceptance of the trial point:** Compute $\rho_k^l = \dfrac{f^l(x_k^l) - f^l(x_k^l + s_k^l)}{t_{q,k}^l(x_k^l, 0) - t_{q,k}^l(x_k^l, s_k^l)}$.
if $\rho_k^l \geq \eta_1$ **then**
$x_{k+1}^l = x_k^l + s_k^l$
else
$x_{k+1}^l = x_k^l$.
end if
• **Regularization parameter update:**
if $\rho_k^l \geq \eta_1$ **then**

$$\lambda_{k+1}^l = \begin{cases} \max\{\lambda_{\min}, \gamma_2 \lambda_k^l\}, & \text{if } \rho_k^l \geq \eta_2, \\ \max\{\lambda_{\min}, \gamma_1 \lambda_k^l\}, & \text{if } \rho_k^l < \eta_2 \end{cases}$$

else
$\lambda_{k+1}^l = \gamma_3 \lambda_k^l$.
end if
$k = k + 1$
end while

Taylor series $T_q(x,s)$, it holds:

$$g(x+s) = T_q(x,s) + \frac{1}{(q-1)!}\int_0^1 (1-\xi)^{q-1}\left(\nabla_x^q g(x+\xi s) - \nabla_x^q g(x)\right)[s]^q\, d\xi, \tag{4.1}$$

$$|g(x+s) - T_q(x,s)| \le \frac{L}{q}\|s\|^{q+1}, \tag{4.2}$$

$$\|\nabla g(x+s) - \nabla_s T_q(x,s)\| \le L\|s\|^q. \tag{4.3}$$

4.1. Global convergence. In this section we prove the global convergence property of the method. Our analysis proceeds in three steps. First, we bound the quantity $|1-\rho_k|$ to prove that λ_k must be bounded above. Then, we relate the norm of the step and the norm of the gradient. Finally, we use these two ingredients to conclude proving that the norm of the gradient goes to zero.

4.1.1. Upper bound for the regularization parameter λ_k. At iteration k we either minimize (decrease) the regularized Taylor model (3.5), or the regularized lower level model (3.6). In both cases, it holds:

$$t_{q,k}^h(x_k^h,0) - t_{q,k}^h(x_k^h,s_k^h) \ge \frac{\lambda_k}{q+1}\|s^l\|^{q+1}, \quad s^l = \begin{cases} s_k^h & \text{if} \quad l = h, \\ s_k^H & \text{if} \quad l = H. \end{cases} \tag{4.4}$$

Let us consider the quantity

$$|1-\rho_k| = \left|1 - \frac{f^h(x_k^h) - f^h(x_k^h+s_k^h)}{t_{q,k}^h(x_k^h,0) - t_{q,k}^h(x_k^h,s_k^h)}\right|. \tag{4.5}$$

If at step k the Taylor model is chosen, from relation (4.2) applied to f^h, (4.4) and (3.8) we obtain the inequality:

$$|1-\rho_k| = \left|\frac{f^h(x_k^h+s_k^h) - T_{q,k}^h(x_k^h,s_k^h)}{T_{q,k}^h(x_k^h,0) - T_{q,k}^h(x_k^h,s_k^h)}\right| \le \frac{L_h(q+1)}{\lambda_k q}.$$

If the lower level model is used, we have from (3.8)

$$|1-\rho_k| = \left|\frac{t_{q,k}^H(x_{0,k}^H,0) - t_{q,k}^H(x_{0,k}^H,s_k^H) - (f^h(x_k^h) - f^h(x_k^h+s_k^h))}{t_{q,k}^H(x_{0,k}^H,0) - t_{q,k}^H(x_{0,k}^H,s_k^H)}\right|.$$

Let us consider the numerator in this expression. From relations (3.3) and (4.1) applied to f^H, using its q-th order Taylor series

$$T_{q,k}^H(x^H,s^H) = f^H(x^H) + \sum_{i=1}^{q}\frac{1}{i!}\nabla_x^i f^H(x^H)\left[s^H\right]^i,$$

it follows

$$
\begin{aligned}
& t^H_{q,k}(x^H_{0,k},0) - t^H_{q,k}(x^H_{0,k},s^H_k) \overset{(3.3)}{=} \\
& T^H_{q,k}(x^H_{0,k},s^H_k) - f^H(x^H_{0,k}+s^H_k) - \sum_{i=1}^{q} \frac{1}{i!}\mathcal{R}(\nabla^i_x f^h(x^h_k))\left[s^H_k\right]^i, \\
\overset{(4.1)}{=} & -\frac{1}{(q-1)!}\int_0^1 (1-\xi)^{q-1}\left[\nabla^q_x f^H(x^H_{0,k}+\xi s^H_k) - \nabla^q_x f^H(x^H_{0,k})\right]\left[s^H_k\right]^q d\xi \\
& -\sum_{i=1}^{q}\frac{1}{i!}\mathcal{R}(\nabla^i_x f^h(x^h_k))\left[s^H_k\right]^i. \qquad (4.6)
\end{aligned}
$$

Similarly the relation (4.1) applied to f^h yields

$$
\begin{aligned}
& f^h(x^h_k) - f^h(x^h_k+s^h_k) = f^h(x^h_k) - T^h_{q,k}(x^h_k,s^h_k) \\
& -\frac{1}{(q-1)!}\int_0^1 (1-\xi)^{q-1}[\nabla^q_x f^h(x^h_k+\xi s^h_k) - \nabla^q_x f^h(x^h_k)]\left[s^h_k\right]^q d\xi. \qquad (4.7)
\end{aligned}
$$

From relation (3.1) we can rewrite $f^h(x^h_k) - T^h_{q,k}(x^h_k,s^h_k)$ as:

$$
f^h(x^h_k) - T^h_{q,k}(x^h_k,s^h_k) = -\sum_{i=1}^{q}\frac{1}{i!}\nabla^i_x f^h(x^h_k)\left[Ps^H_k\right]^i = -\sum_{i=1}^{q}\frac{1}{i!}\mathcal{R}(\nabla^i_x f^h(x^h_k))\left[s^H_k\right]^i.
$$

Then, subtracting (4.7) from (4.6), we obtain

$$
\begin{aligned}
& t^H_{q,k}(x^H_{0,k}) - t^H_{q,k}(x^H_{0,k},s^H_k) - (f^h(x^h_k) - f^h(x^h_k+s^h_k)) = \\
& -\frac{1}{(q-1)!}\int_0^1 (1-\xi)^{q-1}[\nabla^q_x f^H(x^H_{0,k}+\xi s^H_k) - \nabla^q_x f^H(x^H_{0,k})]\left[s^H_k\right]^q d\xi \\
& +\frac{1}{(q-1)!}\int_0^1 (1-\xi)^{q-1}[\nabla^q_x f^h(x^h_k+\xi s^h_k) - \nabla^q_x f^h(x^h_{0,k})]\left[s^H_k\right]^q d\xi.
\end{aligned}
$$

Using Assumption 2, we obtain:

$$
\begin{aligned}
& |t^H_{q,k}(x^H_{0,k},0) - t^H_{q,k}(x^H_{0,k},s^H_k) - (f^h(x^h_k) - f^h(x^h_k+s^h_k))| \\
& \le \frac{1}{(q-1)!}\int_0^1 (1-\xi)^{q-1}|[\nabla^q_x f^H(x^H_{0,k}+\xi s^H_k) - \nabla^q_x f^H(x^H_{0,k})]\left[s^H_k\right]^q| d\xi \\
& +\frac{1}{(q-1)!}\int_0^1 (1-\xi)^{q-1}|[\nabla^q_x f^h(x^h_k+\xi s^h_k) - \nabla^q_x f^h(x^h_k)]\left[s^H_k\right]^q| d\xi \\
& \le \frac{1}{q!}\|s^H_k\|^q \max_{\xi\in[0,1]}\|\nabla^q_x f^H(x^H_k+\xi s^H_k) - \nabla^q_x f^H(x^H_k)\|_{[q]} \\
& +\frac{1}{q!}\|s^h_k\|^q \max_{\xi\in[0,1]}\|\nabla^q_x f^h(x^h_k+\xi s^h_k) - \nabla^q_x f^h(x^h_k)\|_{[q]} \le \frac{1}{q}\left(L_H + L_h\kappa_R^{q+1}\right)\|s^H_k\|^{q+1}.
\end{aligned}
$$

From relation (4.4) we finally obtain:

$$|1-\rho_k| \leq \frac{(q+1)\left(L_H + L_h\kappa_R^{q+1}\right)}{q\lambda_k}.$$

Then, in both cases (when either a Taylor model or a lower level model is used), it exists a strictly positive constant K such that the following relation holds:

$$|1-\rho_k| \leq \frac{K}{\lambda_k}, \qquad K = \begin{cases} \dfrac{(q+1)L_h}{q} & \text{(Taylor model)}, \\ \dfrac{(q+1)\left(L_H + L_h\kappa_R^{q+1}\right)}{q} & \text{(lower level model)}. \end{cases} \tag{4.8}$$

Using this last relation and the updating rule of the regularization parameter, we deduce that λ_k must be bounded above. Indeed, in case of unsuccessful iterations, λ_k is increased. If λ_k is increased, the ratio appearing in the right hand side of (4.8) is progressively decreased, until it becomes smaller than $1-\eta_1$. In this case, $\rho_k > \eta_1$, so a successful step is taken and λ_k is decreased. Hence λ_k cannot be greater than

$$\lambda_{\max} = \frac{K}{1-\eta_1}. \tag{4.9}$$

4.1.2. Relating the steplength to the norm of the gradient. Our next step is to show that the steplength cannot be arbitrarily small, compared to the norm of the gradient of the objective function. If the Taylor model is used, from [6, Lemma 2.3] it follows:

$$\|\nabla_x f^h(x_k^h + s_k^h)\| \leq (L_h + \theta_{\max} + \lambda_{\max})\|s_k^h\|^q := K_1\|s_k^h\|^q. \tag{4.10}$$

If the lower level model is chosen, we have:

$$\begin{aligned} \|R\nabla_x f^h(x_k^h + s_k^h)\| \leq & \left\| R\left[\nabla_x f^h(x_k^h + s_k^h) - \nabla_s T_{q,k}^h(x_k^h, s_k^h)\right] \right\| \\ & + \|R\nabla_s T_{q,k}^h(x_k^h, s_k^h) - \nabla_s t_{q,k}^H(x_{0,k}^H, s_k^H)\| \\ & + \|\nabla_s t_{q,k}^H(x_{0,k}^H, s_k^H) + \lambda_k\|s_k^H\|^{q-1}s_k^H\| + \lambda_k\|s_k^H\|^q. \end{aligned}$$

By (4.3), the first term can be bounded by $\kappa_R L_h\|s_k^h\|^q$. Considering that $s_k^h = Ps_k^H$ and $\|P\| \leq \kappa_R$, we obtain the upper bound $\kappa_R^2 L_h\|s_k^H\|^q$. Regarding the second term, taking into account that from relations $s_k^h = Ps_k^H$, $R = P^T$, and (3.2), for all $p^H \in \mathbb{R}^{n_H}$ it holds:

$$\begin{aligned} \langle \mathcal{R}(\nabla_x^i f^h(x_k^h))\left[s_k^H\right]^{(i-1)}, p^H\rangle &= \langle \nabla_x^i f^h(x_k^h)\left[Ps_k^H\right]^{(i-1)}, Pp^H\rangle \\ &= \langle R\nabla_x^i f^h(x_k^h)\left[Ps_k^H\right]^{(i-1)}, p^H\rangle, \end{aligned}$$

we can write

$$R\nabla_s T_{q,k}^h(x_k^h, Ps_k^H) = \sum_{i=1}^{q}\frac{1}{(i-1)!}R\nabla_x^i f^h(x_k^h)[Ps_k^H]^{(i-1)} = \sum_{i=1}^{q}\frac{1}{(i-1)!}\mathcal{R}(\nabla_x^i f^h(x_k^h))\left[s_k^H\right]^{(i-1)}.$$

Then, from

$$\begin{aligned}\nabla_s t^H_{q,k}(x^H_{0,k}, s^H_k) =& \nabla_x f^H(x^H_{0,k} + s^H_k) \\ &+ \sum_{i=1}^{q} \frac{1}{(i-1)!} \left[\mathcal{R}(\nabla^i_x f^h(x^h_k)) - \nabla^i_x f^H(x^H_{0,k})\right] \left[s^H_k\right]^{(i-1)},\end{aligned} \tag{4.11}$$

we obtain

$$\|R\nabla_s T^h_{q,k}(x^h_k, s^h_k) - \nabla_s t^H_{q,k}(x^H_{0,k}, s^H_k)\| = \left\|\nabla_x f^H(x^H_{0,k} + s^H_k) - \nabla_s T^H_{q,k}(x^H_{0,k}, s^H_k)\right\|,$$

which represents the Taylor remainder for the approximation of $\nabla_x f^H$ by $\nabla_s T^H_{q,k}$. Therefore, by relation (4.3), this quantity can be bounded above by $L_H \|s^H_k\|^q$. The third term, from (3.7), is less than $\theta^H_k \|s^H_k\|^q$. Then, since $\lambda_k \leq \lambda_{\max}$ and $\theta^H_k \leq \theta_{\max}$, we finally obtain

$$\|R\nabla_x f^h(x^h_k + s^h_k)\| \leq \left(\kappa^2_R L_h + L_H + \theta_{\max} + \lambda_{\max}\right) \|s^H_k\|^q := K_2 \|s^H_k\|^q. \tag{4.12}$$

4.1.3. Proof of global convergence. Let us consider the sequence of successful iterations ($\rho_k \geq \eta_1$). They are divided into two groups, $K_{s,f}$ the successful iterations at which the fine model has been employed and $K_{s,l}$ the ones at which the lower level model has been employed. Let us define k_1 the index of the first successful iteration. We remind that at successful iterations $\rho_k \geq \eta_1$. Due to the updating rule of the regularization parameter in Algorithm 3.1 we have $\lambda_k \geq \lambda_{\min}$. Hence from relations (3.8), (4.4), (4.10) and (4.12), (3.4) it follows that:

$$\begin{aligned}f^h(x^h_{k_1}) - \liminf_{k\to\infty} f^h(x^h_k) &\geq \sum_{k\ succ} f^h(x^h_k) - f^h(x^h_k + s^h_k) \\ &\overset{(3.8)}{\geq} \eta_1 \sum_{K_{s,l}} (t^H_{q,k}(x^H_{0,k}) - t^H_{q,k}(x^H_{0,k}, s^H_k)) + \eta_1 \sum_{K_{s,f}} (T^h_{q,k}(x^h_k) - T^h_{q,k}(x^h_k, s^h_k)) \\ &\overset{(4.4)}{\geq} \frac{\eta_1 \lambda_k}{q+1} \left(\sum_{K_{s,l}} \|s^H_k\|^{q+1} + \sum_{K_{s,f}} \|s^h_k\|^{q+1} \right) \\ &\overset{(4.10)+(4.12)}{\geq} \frac{\eta_1 \lambda_{\min}}{q+1} \left(\frac{1}{K_2^{\frac{q+1}{q}}} \sum_{K_{s,l}} \|R\nabla_x f^h(x^h_k + s^h_k)\|^{\frac{q+1}{q}} + \frac{1}{K_1^{\frac{q+1}{q}}} \sum_{K_{s,f}} \|\nabla_x f^h(x^h_k + s^h_k)\|^{\frac{q+1}{q}} \right) \\ &\overset{(3.4)}{\geq} \frac{\eta_1 \lambda_{\min}}{q+1} \left(\frac{1}{K_2^{\frac{q+1}{q}}} \sum_{K_{s,l}} \kappa_H^{\frac{q+1}{q}} \|\nabla_x f^h(x^h_k + s^h_k)\|^{\frac{q+1}{q}} + \frac{1}{K_1^{\frac{q+1}{q}}} \sum_{K_{s,f}} \|\nabla_x f^h(x^h_k + s^h_k)\|^{\frac{q+1}{q}} \right).\end{aligned} \tag{4.13}$$

Hence we conclude that $\sum_{K_{s,f} \cup K_{s,l}} \|\nabla_x f^h(x^h_k + s^h_k)\|$ is a bounded series and therefore has a convergent subsequence. Then, $\|\nabla_x f^h(x^h_k + s^h_k)\|$ converges to zero on the subsequence of successful iterations.

We can then state the global convergence property towards first-order critical points in the following theorem.

Theorem 4.2. *Let Assumptions 1 and 2 hold. Let $\{x^h_k\}$ be the sequence of fine level iterates generated by Algorithm 3.1. Then, $\{\|\nabla_x f^h(x^h_k)\|\}$ converges to zero on the subsequence of successful iterations.*

4.2. Worst-case complexity. We now want to evaluate the worst-case complexity of our methods, to reach a first order stationary point. We assume then that the procedure is stopped as soon as $\|\nabla_x f^h(x_k^h)\| \leq \epsilon$ for $\epsilon > 0$. The proof is similar to that of Theorem 2.5 in [6].

To evaluate the complexity of the proposed methods, we have to bound the number of successful and unsuccessful iterations performed before the stopping condition is met. Let us then define k_f the index of the last iterate for which $\|\nabla_x f^h(x_k^h)\| > \epsilon$, $K_s = \{0 < j \leq k_f \,|\, \rho_j \geq \eta_1\}$ the set of successful iterations before iteration k_f, and K_u its complementary in $\{1, \ldots, k_f\}$. We can use the same reasoning as that used to derive (4.13), but considering in the sum just the successful iterates in K_s. Remind that before termination $\|\nabla_x f^h(x_k^h)\| > \epsilon$ and, in case the lower level model is used, $\|R\nabla_x f^H(x_k^h)\| > \kappa_H \|\nabla_x f^H(x_k^h)\| > \kappa_H \epsilon$ (otherwise at that iteration the Taylor model would have been used). It then follows:

$$f^h(x_{k_1}^h) - \liminf_{k\to\infty} f^h(x_k^h) \geq f^h(x_{k_1}^h) - f^h(x_{k_f+1}^h) = \sum_{j\in K_s} f^h(x_k^h) - f^h(x_k^h + s_k^h)$$
$$\geq \frac{\eta_1 \lambda_{\min}}{q+1} \min\left\{\frac{\kappa_H}{K_2}, \frac{1}{K_1}\right\}^{\frac{q+1}{q}} |K_s| \epsilon^{\frac{q+1}{q}},$$

from which we get the desired bound on the total number of successful iterations. We can then bound the cardinality of K_u, with respect to the cardinality of K_s. From the updating rule of the regularization parameter, it holds:

$$\gamma_1 \lambda_k \leq \lambda_{k+1},\ k \in K_s \qquad \gamma_3 \lambda_k = \lambda_{k+1},\ k \in K_u.$$

Then, proceeding inductively, we conclude that:

$$\lambda_0 \gamma_1^{|K_s|} \gamma_3^{|K_u|} \leq \lambda_{k_f} \leq \lambda_{\max}.$$

Then,

$$|K_s| \log \gamma_1 + |K_u| \log \gamma_3 \leq \log \frac{\lambda_{\max}}{\lambda_0},$$

and, given that $\gamma_1 < 1$, we obtain:

$$|K_u| \leq \frac{1}{\log \gamma_3} \log \frac{\lambda_{\max}}{\lambda_0} + |K_s| \frac{|\log \gamma_1|}{\log \gamma_3}.$$

We can then state the following result.

Theorem 4.3. *Let Assumptions 1 and 2. Let f_{low} denote a lower bound on f and let k_1 denote the index of the first successful iteration in Algorithm 3.1. Then, given an absolute accuracy level $\epsilon > 0$, Algorithm 3.1 needs at most*

$$K_3 \frac{(f(x_{k_1}) - f_{low})}{\epsilon^{\frac{q+1}{q}}} \left(1 + \frac{|\log \gamma_1|}{\log \gamma_3}\right) + \frac{1}{\log \gamma_3} \log\left(\frac{\lambda_{\max}}{\lambda_0}\right)$$

iterations in total to produce an iterate x_k^h such that $\|\nabla_x f(x_k)\| \leq \epsilon$, where

$$K_3 := \frac{q+1}{\eta_1 \lambda_{\min}} \max\left\{K_1, \frac{K_2}{\kappa_H}\right\}^{q+1/q},$$

with K_1 and K_2 defined in (4.10), (4.12), $\gamma_1, \gamma_3, \lambda_0, \lambda_{\min}$ defined in Algorithm 3.1 and $\lambda_{\max}$ defined in (4.9).

Theorem 4.3 reveals that the use of lower level steps does not deteriorate the complexity of the method, and that the complexity bound $O(\epsilon^{-\frac{q+1}{q}})$ is preserved. This is a very satisfactory result, because each iteration of the multilevel methods will be less expensive than one iteration of the corresponding one-level method, thanks to the use of the cheaper lower level models. Consequently, if the number of iterations in the multilevel strategy is not increased, we can expect global computational savings.

4.3. Local convergence. In this section we study the local convergence of the proposed methods towards second-order stationary points. We assume $q \geq 2$ in this section, otherwise the problem is not well-defined. Thanks to the use of high-order models, our methods are expected to attain a fast local convergence rate, especially for growing q. The results reported here are inspired by [44] and extend the analysis proposed therein.

We denote by $\mathcal{X}$ the set of second-order critical points of f, i.e. of points x^* satisfying the second-order necessary conditions:

$$\nabla_x f(x^*) = 0, \quad \nabla_x^2 f(x^*) \succeq 0,$$

i.e. $\nabla_x^2 f(x^*)$ is a symmetric positive semidefinite matrix. We denote by $\mathcal{B}(x, \rho) = \{y \, s.t. \, \|y - x\| \leq \rho\}$ and for all $x \in \mathbb{R}^n$, $\mathcal{L}(f(x)) = \{y \in \mathbb{R}^n \,|\, f(y) \leq f(x)\}$ for $f : \mathbb{R}^n \to \mathbb{R}$.

Remark 4.4. From the assumption that f is a q times continuously differentiable function, it follows that its i-th derivative tensor is locally Lipschitz continuous for all $i \leq q - 1$.

Following [44], we first prove an intermediate lemma that allows us to relate, at generic iteration k, the norm of the step and the distance of the current iterate from the space of second-order stationary points. This lemma holds without need of assuming a stringent non-degeneracy condition, but rather under a local error bound condition, which is a much weaker requirement as it can be satisfied also when f has non isolated second-order critical points.

Assumption 3. *There exist strictly positive scalars $\kappa_{EB}, \rho > 0$ such that*

$$\operatorname{dist}(x, \mathcal{X}) \leq \kappa_{EB} \|\nabla_x f(x)\|, \quad \forall x \in \mathcal{N}(\mathcal{X}, \rho), \tag{4.14}$$

where $\mathcal{X}$ is the set of second-order critical points of f, $\operatorname{dist}(x, \mathcal{X})$ denotes the distance of x to $\mathcal{X}$ and $\mathcal{N}(\mathcal{X}, \rho) = \{x \mid \operatorname{dist}(x, \mathcal{X}) \leq \rho\}$.

This condition has been proposed for the first time in [44]. It is different from other error bound conditions in the literature as, in contrast to them, $\mathcal{X}$ is not the set of first-order critical points, but of second-order-critical points. In addition to being useful for proving convergence, it is also interesting on its own, as it is shown to be equivalent to a quadratic growth condition ([44, Theorem 1]) under mild assumptions on f.

Lemma 4.5. *Let Assumptions 1 and 2 hold. Let $\{x_k^h\}$ be the sequence generated by Algorithm 3.1 and x_k^* be a projection point of x_k^h onto $\mathcal{X}$. Assume that it exists a strictly positive constant $\underline{\rho}$ such that $\{x_k^h\} \in \mathcal{B}(x_k^*, \underline{\rho})$ and that $\nabla_x^2 f$ is Lipschitz continuous in $\mathcal{B}(x_k^*, \underline{\rho})$ with*

Lipschitz constant L_2. Then, it holds:

$$\|s_k^h\| \leq C \operatorname{dist}(x_k^h, \mathcal{X}), \tag{4.15}$$

with

$$C = \begin{cases} C_f = \dfrac{1}{2\lambda_{\max}} \left[L_2 + \sqrt{L_2^2 + 4L_2\lambda_{\max}} \right], & \textit{(Taylor model)}, \\ \kappa_R C_c = \dfrac{\kappa_R}{2\lambda_{\max}} \left[\kappa_R L_2 + \sqrt{\kappa_R^2 L_2^2 + 4L_2\kappa_R\lambda_{\max}} \right], & \textit{(lower level model)}, \end{cases}$$

with $\lambda_{\max}$ and κ_R defined respectively in (4.9) and Assumption 1.

The proof of the lemma is reported in the appendix. This lemma can be used to prove that if it exists an accumulation point of $\{x_k^h\}$ that belongs to $\mathcal{X}$, then the full sequence converges to that point and that the rate of convergence depends on q. First, we can prove that the set of accumulation points is not empty.

Lemma 4.6. *Let Assumptions 1 and 2 hold. Let $\{x_k^h\}$ be the sequence of fine level iterates generated by Algorithm 3.1. If $\mathcal{L}(f(x_k^h))$ is bounded for some $k \geq 0$, then the sequence has an accumulation point that is a first-order stationary point.*

Proof. As $\{f(x_k^h)\}$ is a decreasing sequence, and $\mathcal{L}(f(x_k^h))$ is bounded for some $k \geq 0$, $\{x_k^h\}$ is a bounded sequence and it has an accumulation point. From Theorem 4.2, all the accumulation points are first-order stationary points. ∎

Theorem 4.7. *Let Assumptions 1 and 2 hold. Let $\{x_k^h\}$ be the sequence of fine level iterates generated by Algorithm 3.1. Assume that $\mathcal{L}(f(x_k^h))$ is bounded for some $k \geq 0$ and that it exists an accumulation point x^* such that $x^* \in \mathcal{X}$. Then, the whole sequence $\{x_k^h\}$ converges to x^* and it exist strictly positive constants $c \in \mathbb{R}$ and $\bar{k} \in \mathbb{N}$ such that:*

$$\frac{\|x_{k+1}^h - x^*\|}{\|x_k^h - x^*\|^q} \leq c, \quad \forall k \geq \bar{k}. \tag{4.16}$$

Proof. As x^* is an accumulation point, we have that $\lim_{k\to\infty} \operatorname{dist}(x_k^h, \mathcal{X}) = 0$. Then, it exist ρ and k_1 such that $x_k^h \in \mathcal{N}(\mathcal{X}, \rho)$ for all $k \geq k_1$. Therefore, from Assumption 3 it holds

$$\operatorname{dist}(x_k^h, \mathcal{X}) \leq \kappa_{EB} \|\nabla_x f^h(x_k^h)\|, \quad \forall k \geq k_1. \tag{4.17}$$

Moreover, from Remark 4.4, $\nabla_x^2 f$ is locally Lipschitz continuous, so Lemma 4.5 applies to all $k \geq k_1$.

Let us first consider the case in which the Taylor model is employed. It follows from (4.17), (4.10) and (4.15) that for all $k \geq k_1$

$$\operatorname{dist}(x_{k+1}^h, \mathcal{X}) \leq \kappa_{EB} \|\nabla_x f^h(x_{k+1}^h)\| \leq \kappa_{EB} K_1 \|s_k^h\|^q \leq \kappa_{EB} K_1 C_f^q \operatorname{dist}^q(x_k^h, \mathcal{X}).$$

If the lower level model is employed, from (4.17), (3.4), (4.12) and (A.3) it follows that for all $k \geq k_1$

$$\begin{aligned} \operatorname{dist}(x_{k+1}^h, \mathcal{X}) &\leq \kappa_{EB} \|\nabla_x f^h(x_{k+1}^h)\| \leq \kappa_{EB} \kappa_H \|R \nabla_x f^h(x_{k+1}^h)\| \\ &\leq \kappa_{EB} \kappa_H K_2 \|s_k^H\|^q \leq \kappa_{EB} \kappa_H K_2 C_c^q \operatorname{dist}^q(x_k^h, \mathcal{X}). \end{aligned}$$

Then in both cases, it exists $\bar{C}$ such that

$$\operatorname{dist}(x_{k+1}^h, \mathcal{X}) \leq \bar{C} \operatorname{dist}^q(x_k^h, \mathcal{X}), \quad \forall k \geq k_1,$$

where

$$\bar{C} = \begin{cases} \kappa_{EB} K_1 C_f^q & \text{(Taylor model)}, \\ \kappa_{EB} \kappa_H K_2 C_c^q & \text{(lower level model)}. \end{cases}$$

With this result, we can prove the convergence of $\{x_k^h\}$ with standard arguments. We repeat for example the arguments of the proof of [44, Theorem2] for convenience. Let $\eta > 0$ be an arbitrary value. As $\lim_{k\to\infty} \operatorname{dist}(x_k^h, \mathcal{X}) = 0$, it exists $k_2 \geq 0$ such that

$$\operatorname{dist}(x_k^h, \mathcal{X}) \leq \min\left\{\frac{1}{2\bar{C}}, \frac{\eta}{2C}\right\}, \quad \forall k \geq k_2.$$

Then,

$$\operatorname{dist}(x_{k+1}^h, \mathcal{X}) \leq \bar{C} \operatorname{dist}^q(x_k^h, \mathcal{X}) \leq \frac{1}{2} \operatorname{dist}(x_k^h, \mathcal{X}), \quad \forall k \geq \bar{k} = \max\{k_1, k_2\}.$$

From (4.15), it then holds for all $k \geq \bar{k}$ and $j \geq 0$:

$$\begin{aligned} \|x_{k+j}^h - x_k^h\| &\leq \sum_{i=k}^{\infty} \|x_{i+1}^h - x_i^h\| \leq \sum_{i=k}^{\infty} C \operatorname{dist}(x_i^h, \mathcal{X}) \\ &\leq C \operatorname{dist}(x_k^h, \mathcal{X}) \sum_{i=0}^{\infty} \frac{1}{2^i} \leq 2C \operatorname{dist}(x_k^h, \mathcal{X}) \leq \eta, \end{aligned}$$

i.e. that $\{x_k^h\}_{k\geq\bar{k}}$ is a Cauchy sequence and so the whole sequence is convergent. Finally we establish the q-th order rate of convergence of the sequence. For any $k \geq \bar{k}$,

$$\|x^* - x_{k+1}^h\| = \lim_{j\to\infty} \|x_{k+j+1}^h - x_{k+1}^h\| \leq 2C \operatorname{dist}(x_{k+1}^h, \mathcal{X}) \leq 2C\bar{C} \operatorname{dist}^q(x_k^h, \mathcal{X}). \tag{4.18}$$

Combining this with $\operatorname{dist}(x_k^h, \mathcal{X}) \leq \|x_k^h - x^*\|$, and setting $c = 2C\bar{C}$ we obtain the thesis (4.16).

Therefore $\{x_k^h\}$ converges at least with order q to x^*. ∎

5. Numerical results. In this section, we report on the practical performance of two methods in the family. We have implemented the methods corresponding to $q = 2$ and $q = 3$ in Algorithm 3.1 in Julia [5]. The tests were run on a MacBook Pro 2,4 GHz Intel Core i5 with 4 GB RAM. Note that the method corresponding to $q = 2$ represents a multilevel extension of a version of the well-known adaptive regularization by cubics (that corresponding to AR2 in Algorithm 2.1).

Given $z \in \mathbb{R}^d$, we consider the following nonlinear problem:

$$\begin{cases} -\Delta u(z) + e^{u(z)} = g(z) & \text{in } \Omega, \\ u(z) = 0 & \text{on } \partial\Omega, \end{cases}$$

where g is obtained such that the analytical solution u^* to this problem is known. We consider two instances of this problem:

1. $d = 1$, $u^*(z) = \cos(2\pi z(z-1)) - 1$, $\Omega =]0, 1[$,
2. $d = 2$, $u^*([z_1, z_2]) = \sin(2\pi z_1(1-z_1))\sin(2\pi z_2(1-z_2))$, $\Omega =]0, 1[\times]0, 1[$.

The problem is discretized on a grid of equispaced points z_i, $i = 1, \ldots, n_h^d$. The negative Laplacian operator is discretized using finite difference, giving a symmetric positive definite matrix $A \in \mathbb{R}^{n_h^d \times n_h^d}$, that also takes into account the boundary conditions. The discretized version of the problem is then a system of the form $Au + e^u = g$, where $u, g, e^u \in \mathbb{R}^{n_h^d}$, and their i-th component corresponds to the respective function evaluated in z_i, for $d = 2$ the lexicographical ordering is used for the grid points.

The following nonlinear minimization problem is then solved:

$$\min_{u \in \mathbb{R}^{n_h^d}} \frac{1}{2} u^T A u + \|e^{u/2}\|^2 - g^T u, \tag{5.1}$$

which is equivalent to the system $Au + e^u = g$. The coarse approximations to the objective function arise from a coarser discretization of the problem. Each coarse grid has a dimension that is 2^d times lower than the dimension of the grid on the corresponding upper level.

The prolongation operators P_l from level $l-1$ to l are based on the standard interpolation operator for $d = 1$ and on the nine-point interpolation scheme defined by the stencil $\begin{pmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}$ for $d = 2$. The full weighting operators defined by $R_l = \frac{1}{2^d} P_l^T$ are used as restriction operators [11].

We compare the one-level methods with the proposed multilevel extensions. Parameters common to both methods are set as: $\gamma_1 = 0.85$, $\gamma_2 = 0.5$, $\gamma_3 = 2$, $\lambda_0 = 0.05$, $\eta_1 = 0.1$ and $\eta_2 = 0.75$. For the multilevel procedures we set $\kappa_l = 0.1$.

At each iteration we find an approximate minimizer of the q-th order models (at each level) using a (single level) trust-region approach, as it is done for example in [23, 27], where the recursive trust region [21] is employed. The trust-region solver is stopped according to (3.7) with $\theta_k^l = \|\nabla_x f^l(x_k)\|$ and the trust-region subproblems are solved by the TRS Julia function[1] [1].

The nonlinear process is stopped as soon as $\|\nabla f^{l_{\max}}(x_k^{l_{\max}})\| < \epsilon^{l_{\max}}$, where we have chosen $\epsilon^{l_{\max}} = 10^{-5}$, and $\epsilon^l = \epsilon^{l_{\max}}$ for all l for the multilevel procedures. This stopping criterion is the one commonly used in optimization, and it is convenient from a theoretical point of view to prove the global convergence and complexity results for the family of methods. It is however important to note that the choice of good stopping criterions in multigrid and preconditioned multigrid methods has been the topic of many published works, and the choice of the residual may not always be the best one [2, 10, 13], [34, Ch.11].

We study the effect of the multilevel strategy on the convergence of the method for problems of fixed dimension n_h. We consider the solution of problem (5.1) in case $d = 1$ for $n_h = 256$ and $n_h = 512$ (Table 1) and in case $d = 2$ for $n_h = 32$ and $n_h = 64$ (Table 2), respectively. We allow 4 levels in MARq. We report the results of the average of ten simulations

[1] https://github.com/oxfordcontrol/TRS.jl

Table 1

Solution of the minimization problem (5.1) for $d = 2$, with the one level AR2 method and a four level MAR2. The size of the problems is $n_h^2 = 1024$ and $n_h^2 = 4096$ respectively and the starting guesses are $\bar{u}_1 = 1\,\mathrm{rand}(n_h, 1)$, $\bar{u}_2 = 3\,\mathrm{rand}(n_h, 1)$. The results are the average over ten runs. it_T denotes the number of iterations over ten simulations, it_f the number of iterations in which the fine level model has been used, it_{TR} is the total number of weighted trust-region iterations for the minimization of the model, `RMSE` *the root-mean square error with respect to the true solution, and* `save` *the ratio between the CPU times for AR2 and MAR2.*

		$n_h = 32$		$n_h = 64$	
		AR2	MAR2	AR2	MAR2
$\bar{u}_1$	it_T/it_f	11/11	7/2	23/23	15/4
	it_{TR}	91	43	207	53
	`RMSE`	10^{-3}	10^{-3}	10^{-4}	10^{-4}
	`save`		**2.2**		**4.1**
$\bar{u}_2$	it_T/it_f	27/27	13/4	56/56	22/6
	it_{TR}	220	54	483	79
	`RMSE`	10^{-3}	10^{-3}	10^{-4}	10^{-4}
	`save`		**3.9**		**6.1**

with different random initial guesses of the form $u_0 = a\,\mathrm{rand}(n_h, 1)$, for different values of a. In each simulation the random starting guess is the same for the two considered methods. All the quantities reported in Tables 1 and 2 are the average of the values obtained over ten simulations: it_T denotes the number of total iterations, it_f denotes the number of iterations in which the Taylor model has been used, it_{TR} is the total weighted number of trust-region iterations for the minimization of the model (for MARq trust-region iterations performed at lower levels are weighted by the ratio between the number of variables at the current level over the number of variables at fine level, to take into account their reduced cost), `RMSE` is the root-mean square error with respect to the true solution, `save` is the ratio between the CPU times for ARq and MARq, respectively.

The results reported in Tables 1 and 2 confirm the relevance of MARq as compared to ARq. The numerical experiments highlight that the use of MARq becomes more and more beneficial as the size of the problem increases. It is especially convenient when the initial guess is not so close to the true solution and a higher number of iterations are necessary for the convergence. MARq seems to be much less sensible to the choice of the initial guess than ARq. In all cases, the new multilevel approaches are found to lead to considerable computational savings in terms of CPU time compared to the classical one-level strategies.

6. Conclusions. We have introduced a family of multilevel methods of order $q \geq 1$ for unconstrained minimization. These methods represent an extension of the higher-order methods presented in [6] and of the multilevel trust-region method proposed in [21]. We have proposed a unifying framework to analyse these methods, which is useful to prove their convergence properties and evaluate their worst-case complexity to reach first-order stationary points. As expected, we show that the local rate of convergence and the complexity bound depend on q and high values of q allow both fast local convergence and lower complexity bounds.

We believe this represents a contribution in the optimization field, as the use of multilevel

Table 2

Solution of the minimization problem (5.1) for $d = 1$, with the one level AR3 method and a four level MAR3. The size of the problems is $n_h = 256$ and $n_h = 512$ respectively and the starting guesses are $\bar{u}_1 = 1\,\mathrm{rand}(n_h, 1)$, $\bar{u}_2 = 3\,\mathrm{rand}(n_h, 1)$. The results are the average over ten runs. it_T denotes the number of iterations over ten simulations, it_f the number of iterations in which the fine level model has been used, it_{TR} is the total number of weighted trust-region iterations for the minimization of the model, `RMSE` *the root-mean square error with respect to the true solution, and* `save` *the ratio between the CPU times for AR3 and MAR3.*

		$n_h = 256$		$n_h = 512$	
		AR3	MAR3	AR3	MAR3
$\bar{u}_1$	it_T/it_f	7/7	9/2	18/18	15/2
	it_{TR}	75	45	164	50
	`RMSE`	$5 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	10^{-5}	10^{-5}
	`save`		**2.5**		**4.3**
$\bar{u}_2$	it_T/it_f	23/23	14/1	34/34	20/5
	it_{TR}	215	70	294	74
	`RMSE`	$5 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	10^{-5}	10^{-5}
	`save`		**4.1**		**4.4**

ideas allows to reduce the major cost per iteration of the high-order methods. This gives a first answer to the question posed in [6] about whether the approach presented there can have practical implications, in applications for which computing q derivatives is feasible. Moreover, we have proposed for the first time local convergence results on high-order methods.

We have implemented the multilevel methods corresponding to $q = 2, 3$ and presented numerical results that show the considerable benefits of the multilevel strategy in terms of savings in computational time. Additional numerical results can be found in [12], where the authors apply the multilevel method in the family corresponding to $q = 1$ to problems arising in the training of artificial neural networks for the approximate solution of partial differential equations. This case is particularly insightful as it allows to show the efficiency of multilevel methods even for problems without an underlying geometrical structure.

REFERENCES

[1] S. Adachi, S. Iwata, Y. Nakatsukasa, and A. Takeda, *Solving the trust-region subproblem by a generalized eigenvalue problem*, SIAM J. Opt., 27 (2017), pp. 269–291, https://doi.org/10.1137/16M1058200.

[2] M. Arioli, D. Loghin, and A. Wathen, *Stopping criteria for iterations in finite element methods*, Numer. Math., 99 (2005), pp. 381–410, https://doi.org/https://doi.org/10.1007/s00211-004-0568-z.

[3] L. Badea and R. Krause, *One- and two-level multiplicative Schwarz methods for variational and quasi-variational inequalities of the second kind*, Numer. Math., 120 (2012), p. 573/599, https://doi.org/10.1007/s00211-011-0423-y.

[4] L. Badea, X. Tai, and J. Wang, *Convergence rate analysis of a multiplicative Schwarz method for variational inequalities*, SIAM J. Numer. Anal., 41 (2003), pp. 1052–1073, https://doi.org/10.1137/S0036142901393607.

[5] J. Bezanson, A. Edelman, S. Karpinski, and V. Shah, *Julia: A fresh approach to numerical computing*, SIAM Rev., 59 (2017), pp. 65–98, https://doi.org/10.1137/141000671.

[6] E. G. Birgin, J. L. Gardenghi, J. M. Martínez, S. A. Santos, and P. L. Toint, *Worst-case*

evaluation complexity for unconstrained nonlinear optimization using high-order regularized models, Math. Program., 163 (2017), pp. 359–368, https://doi.org/10.1007/s10107-016-1065-8.

[7] A. Bouaricha, *Tensor methods for large, sparse unconstrained optimization*, SIAM J. Opt., 7 (1997), pp. 732–756, https://doi.org/10.1137/S1052623494267723.

[8] A. Brandt, *A multi-level adaptative solution to boundary-value problems*, Math. Comp., 31 (1977), pp. 333–390, http://www.jstor.org/stable/2006422.

[9] A. Brandt and O. E. Livne, *Multigrid Techniques: 1984 Guide with Applications to Fluid Dynamics*, SIAM, Philadelphia, 2011, https://epubs.siam.org/doi/abs/10.1137/1.9781611970753. Revised Edition.

[10] G. Brethes, O. Allain, and A. Dervieux, *A mesh-adaptive metric-based Full-Multigrid for the Poisson problem*, Meth. Fluids, (2014), https://doi.org/https://doi.org/10.1002/fld.4042.

[11] W. Briggs, V. Henson, and S. McCormick, *A Multigrid Tutorial*, SIAM, Philadelphia, Second ed., 2000, https://doi.org/10.1137/1.9780898719505.

[12] H. Calandra, S. Gratton, E. Riccietti, and X. Vasseur, *On a multilevel Levenberg-Marquardt method for the training of artificial neural networks and its application to the solution of partial differential equations*, eprint arXiv 1904.04685, (2019), https://arxiv.org/abs/1904.04685.

[13] G. Carré and A. Dervieux, *On the application of FMG to variational approximation of flow problems*, Int. J. Comput. Fluid. Dyn., 12 (1999), pp. 99–117, https://doi.org/10.1080/10618569908940817.

[14] C. Cartis, N. Gould, and P. L. Toint, *On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems*, SIAM J. Opt., 20 (2010), pp. 2833–2852, https://epubs.siam.org/doi/abs/10.1137/090774100.

[15] C. Cartis, N. Gould, and P. L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results*, Math. Program., Series A, 127 (2011), pp. 245–295, https://doi.org/10.1007/s10107-009-0286-5.

[16] C. Cartis, N. Gould, and P. L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function- and derivative-evaluation complexity*, Math. Program., 130 (2011), pp. 295–319, https://doi.org/10.1007/s10107-009-0337-y.

[17] X. Chen, P. Toint, and H. Wang, *Complexity of partially separable convexly constrained optimization with non-lipschitzian singularities*, SIAM J. Opt., 29 (2019), pp. 874–903, https://doi.org/10.1137/18M1166511.

[18] A. R. Conn, N. Gould, and P. L. Toint, *Trust region methods*, SIAM, Philadelphia, 2000, https://doi.org/10.1137/1.9780898719857.

[19] P. Deuflhard and M. Weiser, *Global inexact newton multilevel fem for nonlinear elliptic problems*, in Multigrid Methods V, Springer Berlin Heidelberg, 1998, pp. 71–89, https://doi.org/https://doi.org/10.1007/978-3-642-58734-4_4.

[20] J. Dussault, *ARCq: a new adaptive regularization by cubics*, Optim. Methods Softw., 33 (2018), pp. 322–335, https://doi.org/10.1080/10556788.2017.1322080.

[21] S. Gratton, A. Sartenaer, and P. L. Toint, *Recursive trust-region methods for multiscale nonlinear optimization*, SIAM J. Opt., 19 (2008), pp. 414–444, https://doi.org/10.1137/050623012.

[22] A. Griewank, *The modification of Newton's method for unconstrained optimization by bounding cubic terms*, tech. report, Technical report, University of Cambridge, 1981.

[23] C. Gross and R. Krause, *On the convergence of recursive trust-region methods for multiscale nonlinear optimization and applications to nonlinear mechanics*, SIAM J. Numer. Anal., 47 (2009), pp. 3044–3069, https://doi.org/10.1137/08071819X.

[24] C. Gräser, U. Sack, and O. Sander, *Truncated nonsmooth Newton multigrid methods for convex minimization problems*, in Domain Decomposition Methods in Science and Engineering XVIII, no. 70 in Lecture Notes in Computational Science and Engineering, Springer Berlin Heidelberg, 2009, pp. 129–136, http://link.springer.com/chapter/10.1007/978-3-642-02677-5_12.

[25] W. Hackbusch, *Multi-grid methods and applications*, vol. 4 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, Heidelberg, 1985, https://www.springer.com/us/book/9783540127611.

[26] W. Hackbusch and A. Reusken, *Analysis of a damped nonlinear multilevel method*, Numer. Math., 55 (1989), pp. 225–246.

[27] A. Kopaničáková, R. Krause, and R. Tamstorf, *Subdivision-based nonlinear multiscale cloth simu-*

lation, SIAM J. Sci. Comput., 41 (2019), pp. S433–S461, https://doi.org/10.1137/18M1194870.
[28] R. Kornhuber, *Adaptive monotone multigrid methods for nonlinear variational problems*, Teubner–Verlag, 1997, https://www.mi.fu-berlin.de/en/math/groups/ag-numerik/books/Adaptive_monotone_multigrid_methods/index.html.
[29] M. Kočvara and S. Mohammed, *A first-order multigrid method for bound-constrained convex optimization*, Optim. Methods Softw., 31 (2016), pp. 622–644, https://doi.org/10.1080/10556788.2016.1146267.
[30] R. Krause, *A nonsmooth multiscale method for solving frictional two-body contact problems in 2D and 3D with multigrid efficiency*, SIAM J. Sci. Comput., 31 (2009), pp. 1399–1423, https://doi.org/10.1137/070682514.
[31] R. Lewis and S. Nash, *Model problems for the multigrid optimization of systems governed by differential equations*, SIAM J. Sci. Comput., 26 (2005), pp. 1811–1837, https://doi.org/10.1137/S1064827502407792.
[32] R. Lewis and S. Nash, *Using inexact gradients in a multilevel optimization algorithm*, Computational Optimization and Applications, 56 (2013), pp. 39–61, https://doi.org/10.1007/s10589-013-9546-7.
[33] J. Mandel, *A multi-level iterative method for symmetric, positive definite linear complementarity problems*, Appl. Math. Optim., 11 (1984), pp. 77–95, https://doi.org/https://doi.org/10.1007/BF01442171.
[34] J. Mlek and Z. Strakos, *Preconditioning and the Conjugate Gradient Method in the Context of Solving PDEs*, SIAM, Philadelphia, 2014, https://dl.acm.org/doi/book/10.5555/2746448.
[35] S. Nash, *A multigrid approach to discretized optimization problems*, Optim. Methods Softw., 14 (2000), pp. 99–116, https://doi.org/10.1080/10556780008805795.
[36] S. Nash, *Properties of a class of multilevel optimization algorithms for equality constrained problems*, Optim. Methods Softw., 29 (2014), pp. 137–159, https://doi.org/10.1080/10556788.2012.759571.
[37] Y. Nesterov and B. Polyak, *Cubic regularization of Newton method and its global performance*, Math. Program., (2006), pp. 177–205, https://link.springer.com/article/10.1007/s10107-006-0706-8.
[38] J. Nocedal and S. Wright, *Numerical Optimization*, Springer-Verlag, New York, 2006, https://doi.org/10.1007/978-0-387-40065-5.
[39] X. Tai and J. Xu, *Global and uniform convergence of subspace correction methods for some convex optimization problems*, Math. Comput., 71 (2002), pp. 105–124, https://doi.org/https://doi.org/10.1090/S0025-5718-01-01311-4.
[40] P. L. Toint, *Nonlinear stepsize control, trust regions and regularizations for unconstrained optimization*, Optim. Methods Softw., 28 (2013), pp. 82–95, https://www.tandfonline.com/doi/abs/10.1080/10556788.2011.610458?journalCode=goms20.
[41] S. Wang and S. Liu, *A tensor trust-region model for nonlinear system*, J. Inequal. Appl., 2018 (2018), p. 343, https://doi.org/10.1186/s13660-018-1935-0.
[42] Z. Wen and D. Goldfarb, *A line search multigrid method for large-scale nonlinear optimization*, SIAM J. Opt., 20 (2009), pp. 1478–1503, https://doi.org/10.1137/08071524X.
[43] Y. Yuan, *Recent advances in trust region algorithms*, Math. Program., 151 (2015), pp. 249–281, https://link.springer.com/article/10.1007/s10107-015-0893-2.
[44] M. Yue, Z. Zhou, and A. Man-Cho So, *On the quadratic convergence of the cubic regularization method under a local error bound condition*, SIAM J. Opt., 29 (2019), pp. 904–932, https://doi.org/10.1137/18M1167498.

Appendix A. Proof of Lemma 4.5 .

In this appendix we report the proof of Lemma 4.5. We restate it here for convenience of the reader.

Lemma 4.5. *Let Assumptions 1 and 2 hold. Let $\{x_k^h\}$ be the sequence generated by Algorithm 3.1 and x_k^* be a projection point of x_k^h onto $\mathcal{X}$. Assume that it exists a strictly positive constant $\underline{\rho}$ such that $\{x_k^h\} \in \mathcal{B}(x_k^*, \underline{\rho})$ and that $\nabla_x^2 f$ is Lipschitz continuous in $\mathcal{B}(x_k^*, \underline{\rho})$ with Lipschitz constant L_2. Then, it holds:*

$$\|s_k^h\| \leq C \operatorname{dist}(x_k^h, \mathcal{X}),$$

with

$$C = \begin{cases} C_f = \dfrac{1}{2\lambda_{\max}}\left[L_2 + \sqrt{L_2^2 + 4L_2\lambda_{\max}}\right], & \textit{(Taylor model)},\\ \kappa_R C_c = \dfrac{\kappa_R}{2\lambda_{\max}}\left[\kappa_R L_2 + \sqrt{\kappa_R^2 L_2^2 + 4L_2\kappa_R\lambda_{\max}}\right], & \textit{(lower level model)}, \end{cases}$$

with $\lambda_{\max}$ *and* κ_R *defined respectively in* (4.9) *and Assumption 1.*

Proof. The proof is divided into two parts. We first consider the case in which s_k^h has been obtained from the approximate minimization of the Taylor model, and then the case in which it has been obtained as prolongation of the step obtained from the approximate minimization of the coarse model.

Le us then assume that the Taylor model has been employed. Reminding that $\nabla_x f^h(x^*) = 0$ for each $x^* \in \mathcal{X}$, and definition (3.5) we obtain:

$$\begin{aligned} \nabla_s m_{q,k}^h(x_k^h, s_k^h; \lambda_k) = &- \nabla_x f^h(x_k^*) + \nabla_x f^h(x_k^h) + \nabla_x^2 f^h(x_k^h) s_k^h \\ &+ H(s_k^h) + \lambda_k \|s_k^h\|^{q-1} s_k^h, \end{aligned} \tag{A.1}$$

with

$$H(s_k^h) = \sum_{i=3}^{q} \frac{1}{(i-1)!} \nabla_x^i f^h(x_k^h)[s_k^h]^{(i-1)}.$$

Some algebraic manipulations (adding $\nabla_x^2 f^h(x_k^*)(x_{k+1}^h - x_k^*)$ to both sides of (A.1) and expressing $(x_{k+1}^h - x_k^*) = s_k^h + (x_k^h - x_k^*)$) lead to:

$$\begin{aligned} &\left(\nabla_x^2 f^h(x_k^*) + \lambda_k \|s_k^h\|^{q-1}\right)(x_{k+1}^h - x_k^*) = \\ &\nabla_s m_{q,k}^h(x_k^h, s_k^h; \lambda_k) + \nabla_x f^h(x_k^*) - \nabla_x f^h(x_k^h) - \nabla_x^2 f^h(x_k^*)(x_k^* - x_k^h) \\ &- H(s_k^h) + (\nabla_x^2 f^h(x_k^*) - \nabla_x^2 f^h(x_k^h)) s_k^h - \lambda_k \|s_k^h\|^{q-1}(x_k^* - x_k^h). \end{aligned}$$

Using the fact that $\nabla_x^2 f^h(x_k^*) \succeq 0$, the stopping criterion (3.7), and the triangle inequality, it follows

$$\begin{aligned} \lambda_k \|s_k^h\|^{q-1} \|x_{k+1}^h - x_k^*\| \leq &\theta_k^h \|s_k^h\|^q + \|\nabla_x f^h(x_k^*) - \nabla_x f^h(x_k^h) - \nabla_x^2 f^h(x_k^*)(x_k^* - x_k^h)\| + \\ &\|H(s_k^h)\| + \|\nabla_x^2 f^h(x_k^*) - \nabla_x^2 f^h(x_k^h)\| \|s_k^h\| + \lambda_k \|s_k^h\|^{q-1} \|x_k^* - x_k^h\|. \end{aligned}$$

Using the Lipschitz continuity of $\nabla_x^2 f$ in $\mathcal{B}(x_k^*, \rho)$, the relation (4.3) with $q = 2$ and the triangle inequality $\|x_{k+1}^h - x_k^*\| \geq \|x_{k+1}^h - x_k^h\| - \|x_k^h - x_k^*\| = \|s_k^h\| - \|x_k^h - x_k^*\|$, we obtain:

$$\lambda_k \|s_k^h\|^q \leq \theta_k^h \|s_k^h\|^q + L_2 \|x_k^h - x_k^*\|^2 + \|H(s_k^h)\| + L_2 \|x_k^* - x_k^h\| \|s_k^h\| + 2\lambda_k \|s_k^h\|^{q-1} \|x_k^* - x_k^h\|.$$

Notice that

$$\begin{aligned} &\theta_k^h \|s_k^h\|^q + L_2 \|x_k^h - x_k^*\|^2 + \|H(s_k^h)\| + L_2 \|x_k^* - x_k^h\| \|s_k^h\| + 2\lambda_k \|s_k^h\|^{q-1} \|x_k^* - x_k^h\| \\ &\geq L_2 \|x_k^h - x_k^*\|^2 + L_2 \|x_k^* - x_k^h\| \|s_k^h\|. \end{aligned}$$

We can then study when the inequality holds

$$\lambda_k \|s_k^h\|^q \leq L_2 \|x_k^h - x_k^*\|^2 + L_2 \|x_k^* - x_k^h\| \|s_k^h\|.$$

The right hand side of the inequality is expressed as a polynomial of $\|s_k^h\|$ of order 1 with positive value in 0, so the inequality will be true if $\|s_k^h\|$ is small enough. We can then assume $\|s_k^h\| < 1$, so that $\|s_k^h\|^q \leq \|s_k^h\|^2$ if $q \geq 2$. Then, we have that

$$\lambda_k \|s_k^h\|^q \leq \lambda_k \|s_k^h\|^2.$$

We can then solve

$$L_2 \|x_k^h - x_k^*\|^2 + L_2 \|x_k^* - x_k^h\| \|s_k^h\| - \lambda_k \|s_k^h\|^2 \geq 0.$$

The solution leads to

$$\|s_k^h\| \leq C_f \|x_k^h - x_k^*\|, \quad C_f = \frac{1}{2\lambda_k} \left[L_2 + \sqrt{L_2^2 + 4L_2\lambda_k} \right]. \tag{A.2}$$

Let us now consider the case in which the lower level model is used. The idea is similar as in the previous case. Reminding (4.11) and that $R\nabla_x f^h(x_k^*) = 0$, we have:

$$\begin{aligned} \nabla_s m_k^H(x_{0,k}^H, s_k^H; \lambda_k) =& \nabla_x f^H(x_{0,k}^H + s_k^H) - \nabla_s T_{q,k}^H(x_{0,k}^H, s_k^H) + R\nabla_x f^h(x_k^*) + \\ & \sum_{i=1}^{q} \frac{1}{(i-1)!} \mathcal{R}(\nabla_x^i f^h(x_k^h)) \left[s_k^H\right]^{(i-1)} + \lambda_k \|s_k^H\|^{q-1} s_k^H. \end{aligned}$$

Algebraic manipulations (adding $R\nabla_x^2 f^h(x_k^*)(x_{k+1}^h - x_k^*)$ to both sides and expressing $(x_{k+1}^h - x_k^*) = s_k^h + (x_k^h - x_k^*)$) lead to:

$$\begin{aligned} R\nabla_x^2 f^h(x_k^*)(x_{k+1}^h - x_k^*) =& \nabla_s m_k^H(x_{0,k}^H, s_k^H; \lambda_k) - \nabla_x f^H(x_{0,k}^H + s_k^H) \\ & + \nabla_s T_{q,k}^H(x_{0,k}^H, s_k^H) + R\nabla_x f^h(x_k^*) - R\nabla_x f^h(x_k^h) - R\nabla_x^2 f^h(x_k^*)(x_k^* - x_k^h) \\ & - H_H(s_k^H) + R(\nabla_x^2 f^h(x_k^*) - \nabla_x^2 f^h(x_k^h)) s_k^h - \lambda_k \|s_k^H\|^{q-1} s_k^H, \end{aligned}$$

where

$$H_H(s_k^H) = \sum_{i=3}^{q} \frac{1}{(i-1)!} \mathcal{R}(\nabla_x^i f^h(x_k^H)) \left[s_k^H\right]^{(i-1)}.$$

Further, we can write

$$\begin{aligned} R\nabla_x^2 f^h(x_k^*)(x_{k+1}^h - x_k^*) =& R\nabla_x^2 f^h(x_k^*)(x_{k+1}^h - x_k^h) + R\nabla_x^2 f^h(x_k^*)(x_k^h - x_k^*) \\ =& R\nabla_x^2 f^h(x_k^*) P s_k^H + R\nabla_x^2 f^h(x_k^*)(x_k^h - x_k^*). \end{aligned}$$

Then

$$\begin{aligned} (R\nabla_x^2 f^h(x_k^*) P + \lambda_k \|s_k^H\|^{q-1}) s_k^H =& \nabla_s m_{q,k}^H(x_k^H, s_k^H; \lambda_k) - \nabla_x f^H(x_{0,k}^H + s_k^H) \\ & + \nabla_s T_{q,k}^H(x_{0,k}^H, s_k^H) + R\nabla_x f^h(x_k^*) - R\nabla_x f^h(x_k^h) - R\nabla_x^2 f^h(x_k^*)(x_k^* - x_k^h) \\ & - H_H(s_k^H) + R(\nabla_x^2 f^h(x_k^*) - \nabla_x^2 f^h(x_k^h)) P s_k^H - R\nabla_x^2 f^h(x_k^*)(x_k^h - x_k^*). \end{aligned}$$

We can again use relation (4.3) (applied to $f^H, T^H_{q,k}$ with constant L_H and to $f, T^h_{2,k}$ with constant L_2), (3.7), the fact that $R\nabla^2_x f^h(x^*_k)P$ is still positive definite, and Assumption 1 together with relation $s^h_k = Ps^H_k$, to deduce that:

$$\begin{aligned}\lambda_k\|s^H_k\|^q \leq (\theta^H_k + L_H)\|s^H_k\|^q + \kappa_R L_2\|x^*_k - x^h_k\|^2 + \|H_H(s^H_k)\| \\ + \kappa^2_R L_2\|x^*_k - x^h_k\|\|s^H_k\| + \|R\nabla^2_x f^h(x^*_k)(x^h_k - x^*_k)\|.\end{aligned}$$

We remark that

$$\begin{aligned}(\theta^H_k + L_H)\|s^H_k\|^q + \kappa_R L_2\|x^*_k - x^h_k\|^2 + \|H_H(s^H_k)\| + \kappa^2_R L_2\|x^*_k - x^h_k\|\|s^H_k\| \\ +\|R\nabla^2_x f^h(x^*_k)(x^h_k - x^*_k)\| \geq \kappa_R L_2\|x^*_k - x^h_k\|^2 + \kappa_R L_2\|x^*_k - x^h_k\|\|s^H_k\|.\end{aligned}$$

As previously, we can solve the following inequality:

$$\lambda_k\|s^H_k\|^2 \leq \kappa_R L_2\|x^*_k - x^h_k\|^2 + \kappa^2_R L_2\|x^*_k - x^h_k\|\|s^H_k\|,$$

and conclude that:

$$\|s^H_k\| \leq C_c\|x^h_k - x^*_k\|, \quad C_c = \frac{\left[\kappa_R L_2 + \sqrt{\kappa^2_R L^2_2 + 4L_2\kappa_R\lambda_k}\right]}{2\lambda_k}. \tag{A.3}$$

We can then use the fact that $\lambda_k \leq \lambda_{\max}$ for all k and that $\|s^h_k\| \leq \kappa_R\|s^H_k\|$ to conclude that in all cases it exists a constant C such that $\|s^h_k\| \leq C\|x^h_k - x^*_k\|$. ∎